Developing a Benchmark for Pronunciation Feedback: Creation of a Phonemically Annotated Speech Corpus of isiZulu Language Learner Speech

Alexandra O'Neil,[1] **Nils Hjortnæs,**[1] **Zinhle Nkosi,**[2] **Thulile Ndlovu,**[2]
Zanele Mlondo,[2] **Ngami Phumzile Pewa,**[2] **and Francis Tyers**[1]
[1]Indiana University, Bloomington, IN USA
[2]University of KwaZulu-Natal, Durban, South Africa
{aconeil, nhjortn, ftyers}@iu.edu, {nkosiz, ndlovut, mlondoz1, pewan}@ukzn.ac.za

Abstract

Pronunciation of the phonemic inventory of a new language often presents difficulties to second language (L2) learners. These challenges can be alleviated by the development of pronunciation feedback tools that take speech input from learners and return information about errors in the utterance. This paper presents the development of a corpus designed for use in pronunciation feedback research. The corpus is comprised of gold standard recordings from isiZulu teachers and recordings from isiZulu L2 learners that have been annotated for pronunciation errors. Exploring the potential benefits of word-level versus phoneme-level feedback necessitates a speech corpus that has been annotated for errors on the phoneme-level. To aid in this discussion, this corpus of isiZulu L2 speech has been annotated for phoneme-errors in utterances, as well as suprasegmental errors in tone.

Keywords: speech corpus, language learning, pronunciation feedback, less-resourced languages

1. Introduction

Pronunciation of the phonemic inventory of a new language often presents difficulties to second language (L2) learners. To address this, language learning tools, such as DuoLingo[1], ELSA[2], Memrise[3], etc. allow users to practice pronunciation. These are all proprietary. For research to progress in this area, annotated corpora are needed.

While existing corpora can be stretched to work for various projects, they often do not apply well to research in computer-assisted pronunciation training (CAPT). Research in CAPT aims to improve a speaker's pronunciation by providing corrective feedback on their speech, often through contrasting their production with that of a native speaker (Nazir et al., 2023). The specificity of this type of research question motivates a corpus that is representative of the production of language learner speech, including learner errors. To bridge this gap in available corpora for CAPT research, we detail the creation of a corpus of isiZulu language learner speech which includes recordings from students and teachers, as well as annotation of errors in student recordings.

This corpus is specifically designed to assist in evaluating the performance of pronunciation feedback tools for second language learning which are used to give language learners feedback on their utterances. While feedback indicating the erroneous phoneme has shown to be effective in language learning (Engwall, 2012), Phan et al. (2023) note that the availability of corpora containing L2 learner speech is the main impediment for producing phoneme-level speech for less-resourced languages. In addition to being useful for computational tools, Lozano and Mendikoetxea (2013) not the utility of second-language learner speech corpora in the field of second language acquisition.

Recordings and annotation in our corpus come from speakers and learners of the Zulu language, isiZulu. IsiZulu was chosen due to its large phonemic inventory of consonants, including phonemic aspiration contrast for stops, implosives, and 15 distinct clicks (Canonici, 1989; Doke, 1961). A larger phonemic inventory requires the recognition of more distinctions in the input, so testing on such a corpus necessitates a model with robust ability to distinguish between similar phonemes. Starting with pronunciation feedback for a phonemically-rich language facilitates future research on languages where some of the phonemic categories are condensed, as generalization is easier to add than discernment. Additionally, isiZulu provides a case study for how the design works for a lesser-resourced language.

2. Prior Work

Limited research exists for the topic of mispronunciation detection as existing work in Second Lan-

[1] https://www.duolingo.com/
[2] https://elsaspeak.com/en/
[3] https://www.memrise.com/

guage Acquisition (SLA) prioritizes written corpora (Granger, 2011) and the use of Automatic Speech Recognition (ASR) for Computer-Assisted Pronunciation Training has only recently become a topic of increased interest as a result of improvements in ASR.

Second Language Acquisition The motivation for creating a corpus of language learner speech is based in the desire to benchmark models for phoneme-based pronunciation feedback, an idea that comes from the field of SLA. Williams (1979) notes that the learner's perceptual understanding of a language is tied to the speaker's production of the language. Feedback is able to assist in language learning as it makes the speaker aware of a difference in their pronunciation and the target pronunciation (Schmidt, 1990). Noticing these differences changes the speaker's perceptual understanding of the language and subsequently their production. Prior work from SLA was also considered in the collection of speaker information, organization of the data, the accessibility of the resources and the selection of sentence, in accordance with best practices as described in (MacWhinney, 2017; Reppen, 2022; Baden et al., 2022).

Automatic Speech Recognition There are several speech corpora for isiZulu, of which we are aware of 3 that are specifically intended for speech recognition. The largest is the NCHLT isiZulu Speech Corpus, which contains approximately 56 hours of data (Barnard et al., 2014). Additional, smaller corpora include CatchWord Language and Speech Technologies'[4] and the multilingual Lwazi isiZulu ASR corpus[5] (Barnard et al., 2009). Our corpus is comparable in size to these latter corpora, though attention should be paid by users to the difference in contents as our corpus intentionally includes non-native speakers.

Historically, Hidden Markov Model (HMM) based speech recognition has shown a great deal of success (Malik et al., 2021). This success is reflected in prior work on automatic pronunciation scoring using HMMs (Franco et al., 2010; Dalby and Kewley-Port, 1999). For this purpose, a corpus was collected by Bratt et al. (1998), but is no longer available. Coulange (2023) notes that many modern systems probably use neural network based models, as those have become more popular in recent years for speech recognition (Malik et al., 2021), but specifics and corpora used are not publicly available. Additionally, the most recent attempts at using ASR to support mispronunciation detection have utilized open models, such as wav2vec2.0 and Whisper, for the task (Peng et al., 2021; Phan et al., 2023; Li et al., 2023).

ID	L1	Gender	Age	Semesters
1	English	F	19	3
2	English	F	19	3
3	English	F	20	3
4	siSwati	F	18	1
5	siSwati	F	19	1
6	isiXhosa	M	21	1
7	isiXhosa	M	21	1
8	English	F	19	3
9	English	F	18	1
10	English	M	19	1
11	English	F	19	3
12	English	M	19	3
101	isiZulu	F	45	–
102	isiZulu	F	33	–
103	isiZulu	M	30	–

Table 1: Demographics of the Participants. IDs greater than 100 are teachers who provided recordings. The semester column indicates the number of semesters studied so far.

3. Corpus Contents

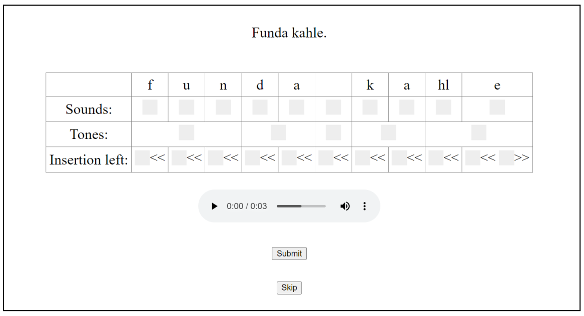
The corpus consists of 12,065 recordings stored as wav files spanning 12.2 hours; 9,626 from students and 2,439 from teachers. The recordings are labelled with a unique speaker ID, age, gender, other languages, pre-university experience with isiZulu, semester(s) of study, birthplace, and place of residence. Of the 9,626 student clips, 5,539 sentences have rankings from one teacher, 3,162 sentences have rankings from 2 teachers and 925 sentences have rankings from all 3 teachers. All clips have at least 1 set of annotations.

Participants The participants recorded for this corpus are from the University of KwaZulu-Natal, located in Durban, South Africa. An overview of the demographic information is provided in Table 1. 12 students provided recordings, of which 8 were female and 4 were male. The students ranged from 18 to 21 years old, with half having completed 1 semester of isiZulu study and the others having completed 3 semesters. 8 students were native speakers of English, while 2 were native speakers of isiXhosa, and 2 were native speakers of siSwati. The sample of the corpus was restricted by interest in participation and availability of students. The distribution of English and isiXhosa speakers is comparable to the demographic constituency of the KwaZulu-Natal province (Ndebele

[4] `CatchWordLanguageandSpeechTechnologies`

[5] `https://repo.sadilar.org/handle/20.500.12185/463`

and Zulu, 2017), but the sample does not comprehensively cover second-language learners languages of all backgrounds of isiZulu. 2 female teachers and 1 male teacher provided recordings. Two of the teachers teach isiZulu as a second language while 1 of the teachers teaches mother-tongue isiZulu. Those who provided recordings are considered our participants. Teachers who marked errors in student recordings are referred to as annotators throughout the paper.



Funda kahle.

	f	u	n	d	a		k	a	hl	e
Sounds:	□	□	□	□	□	□	□	□	□	□
Tones:	□			□		□	□		□	
Insertion left:	□<<	□<<	□<<	□<<	□<<	□<<	□<<	□<<	□<<	□<< □>>

Figure 1: An example of the interface teachers saw when providing feedback on a sentence. The display sentence *Funda kahle.* translates to the command "Study well."

Annotation Layers Annotators were tasked with marking mispronunciations of phonemes and tone, as well as sound insertions. Annotations were collected using a simple web interface[6], see Figure 1. In order to mark these errors, the annotation interface had three tiers: "sound", "tone", and "insertion left." For presentation to teachers, the sentences remained in isiZulu orthography, as knowledge of the International Phonetic Alphabet is not assumed. Although they are presented in the orthography of the language, both the character segmentation and tonal boundaries are determined by the phonemic and syllabification rules of the language. For example, the characters "hl" correspond to one sound, [ɬ] and thus one box in the sounds tier. Prior to annotation, we confirmed that the segmentation of characters is logical to how the annotators think about the sounds in isiZulu.

Syllabification is done with preference to CV structure, as proposed by (Khumalo, 1984), although a CVN structure is noticed when there isn't a single corresponding sound in isiZulu. For example, "ngq" represents one phoneme, the voiced nasalized postalveolar click [!̬̃], so the "n" is never separated from the rest of the articulation. However, "nd" is broken into the phoneme [n] and [d] since these are two separate sounds in the language and not a co-articulation.

Mispronunciation Detection While the ultimate goal of the corpus is to aid in pronunciation feedback research, analysis of the corpus itself provides insight on its ability to represent the speech of language learners. Our analysis compares the reported errors to those predicted and studied in Second Language Acquisition (SLA) literature.

For brevity our analysis of the reported mispronunciations focuses on the phonemic level, with a more thorough discussion of tonal error left to future research. Figure 2 shows the total errors in the corpus normalized by the total number of occurrences of the phoneme in the corpus. A preliminary analysis of reported phonemic errors from the feedback of the teachers corresponds to existing research in SLA, particularly regarding expectations for isiZulu learners and the production of voiceless unaspirated stops. The various clicks also tend to cause learners trouble, as can be seen in the orange bars.

Best et al. (2001) studies the role of perception for second language learners of isiZulu and finds that common errors in perception occur when a student fails to notice the distinction between two phonemes and maps both phonemes to one phonemic category in their native language. Major (2008) describes the tendency of native speakers of English to produce French [t̪] as [tʰ]. As Figure 2 shows, many errors are noted on the consonants [p], [t], and [k]. Looking at existing studies of isiZulu students and considering that the native language of more than 2/3 of the speakers is English, the errors on these consonants can be attributed to a lack of distinction in the perception of plain versus aspirated stops in isiZulu and a transfer of aspiration preference from English. While there are many more interesting observations from the annotated mispronunciations, we leave these to future SLA research.

Inter-Annotator Agreement Inter-Annotator Agreement (IAA) is calculated using Krippendorff's alpha (Krippendorff, 2011), it allows for data collected from more than 2 annotators and accounts for sentences that have at least 2 annotations. The 9,626 sentences of the corpus contain 169,074 phonemes, judged as correct or incorrect, for a total of 256,831 phoneme judgements across annotators. Judgements at the phoneme-level are used to calculate IAA. Of these 169,074 phonemes, 71,370 have judgements from 2 or more annotators and can be used to calculate the IAA. Using the Pypi implementation of Krippendorff's alpha (Castro, 2017), IAA is .585, which is considered moderate agreement (Landis and Koch, 1977), although an IAA level of .8 or above is recommended for reliable agreement (Artstein and Poesio, 2008). For the purpose

[6] the code for this interface can be found at https://github.com/hjortnaes/pronunciation_feedback_survey

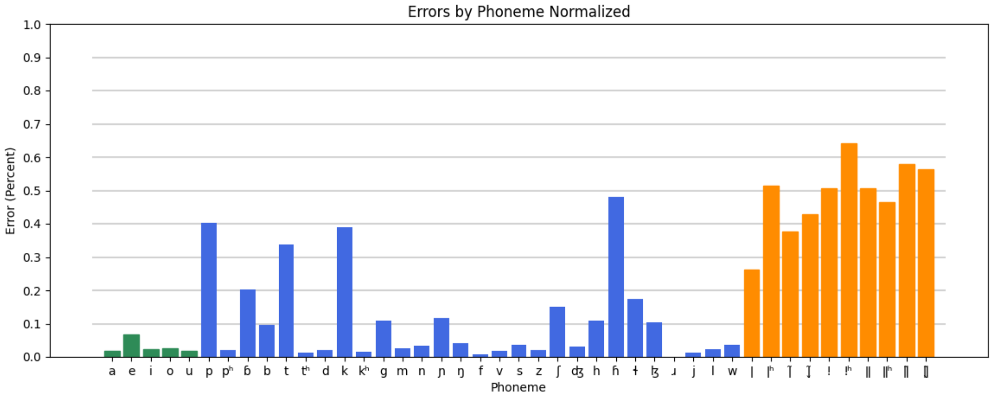


Figure 2: The above graph depicts mispronunciations normalized by total occurrences in the corpus. The x-axis is organized by vowels (in green), non-click consonants (in blue) sorted by place and manner of articulation, and click consonants (in orange) sorted by place and manner of articulation.

of this study, .585 demonstrates satisfactory agreement, as the task of deciding the degree to which mispronunciation of a phoneme spreads to surrounding phonemes is expected to be challenging for annotators and result in some spread of errors from phoneme-to-phoneme, with the trade-off for fine-grained error annotation being realized in IAA.

4. Methodology

Our methodology contains 5 steps: sentence selection, recording, processing, organization, and annotation. As annotation is discussed as part of section 3, we do not repeat it here.

Sentence Selection As the target use case for this corpus is language learning tools, sentences were extracted from language learning materials. This corpus is based on a Zulu language learning textbook so as to use sentences authentic to a language learning environment. Because textbooks increase in difficulty unit-by-unit, the difficulty of the sentences in the corpus can be filtered using the unit number as a key.

To build an open-source corpus, the sentences were extracted from a non-copyrighted source, so we selected a textbook published in 1921 that had entered the public domain, *“Elementary Zulu: A Course of Elementary Lessons in the Zulu Language: Intended Chiefly for Beginners and Junior”* (W, 1921). The orthography of the textbook was outdated, but the discrepancies between older isiZulu orthography and modern orthography are consistent, so regular expressions were used to update to the current orthography. For example, the textbook used an isolating orthography, such that the phrase *Ngiyamthanda.* “I like him/her.” would be written *Ngi ya m thanda*. Using knowledge of isiZulu verbal construction all occurrences of a personal subject concord, the lengthening marker “ya,” and the 3rd person object concord could be fixed by querying the text for the regular expression

```
(ngi|u|si|ni|ba)\sya\s((m)\s)?
```

and replacing it with

```
\1ya\3
```

Following the conversion of the orthography and removal of duplicate sentences, 803 sentences from the textbook were reviewed in consultation with a native speaker. Any errors resultant of the OCR or missed by regular expressions were addressed in this step. Additionally, versions of words that occurred twice, but are less common in modern speech were replaced with the common term, such as replacing multiple occurrences of the root *ukuloba* “to write (archaic)” with *ukubhala* “to write.” These terms were identified by the native speaker. An isiZulu teacher added an additional twenty sentences to the list to reflect sentences that are commonly taught in classes, but were absent from the textbook, such as *Uneminyaka emingaki?* “How old are you?” and *Ngiyajabula ukukwazi.* “I’m happy to meet you.” An additional eight sentences were added to the list that included rarer phonemes missing from the sentences in the list. Each sentence in the list is labelled with the unit it first appeared in or, in the case of the added sentences with codes marking the reason for addition, respectively PHREX and PHON.

Recording The open-source audio software Audacity was used to record, process, and export the recordings from participants. The sentence list was then randomized[7] and split into blocks of 50 for elicitation. Details on the settings used for elicitation can be found in Appendix A.

Internal microphones on the researchers' laptops were used to collect recordings. External microphones were tested, and were found to produce a lower quality recording. The initial recordings used the default sound settings on the laptops, but the built-in noise dampening software dampened the production of clicks in the recordings. This example of unintended altering of a phoneme in the language highlights the importance of turning off all default audio processing in the tools used for recording.

Recordings were collected in an empty classroom on the university campus. Students were given the list at the time of recording and were encouraged to sight read the sentences. False starts and mispronunciations were maintained in the recordings, as these are also expected in regular speech. A sentence was only re-recorded in cases of extreme background noise or clipping of the audio.

Processing In order to better maintain an authentic recording environment and ensure no accidental processing out of clicks, the only processing step performed was normalization. In addition to providing more uniformity for the annotation process done by teachers, Ibrahim et al. (2017) encourage this step by saying "In this way the peak energy value of each word is zero decibels and the recognition system is relatively insensitive to the difference in gain between different recordings."

Organization The organization step starts with the export process from Audacity. Once normalized, the clips can be exported as .wav files using the "Export multiple" option and checking the option to use the track names as file names. Once the export is finished, Audacity generates a list of the exported track titles which can be used to rename the files en mass. The recordings are named by elicited phrase, sentence difficulty, and speaker ID. For this corpus, these categories correspond to elicitation order number, the source book chapter, and speaker ID. The index in the randomized list and source location code are consistently mapped to one another, since the elicitation order was the same for every speaker. For example, the recording labelled 10-14-001.wav corresponds to the tenth elicited sentence which is from chapter 14 and has been recorded by the speaker with ID 001.

5. Conclusion

This research proposes and demonstrates a methodology for the creation of second-language learner speech corpora. The goal of this corpus design is to provide a resource for realistic, data-grounded evaluation of current mispronunciation feedback systems. The resulting corpus can be found on the South African Centre for Digital Language Resources (SADiLaR) website[8]. This corpus fills a needed niche for computational linguistics, specifically focusing on language learning. The data can also be used for Automatic Speech Recognition, contributing to the total annotated speech data available for isiZulu. The corpus shows that the most difficult phonemes for the learners of isiZulu are the unaspirated stops and clicks. Finally, and most importantly, while there is disagreement among our annotators, there is enough agreement to evaluate language learning feedback systems using this corpus. Future work can use this corpus to evaluate the effectiveness of pronunciation feedback systems.

6. Acknowledgements

This project was funded by Indiana University Research through the Social Sciences Research Funding Program. We would also like to thank all of the professors, staff, and students at the University of KwaZulu-Natal Edgewood campus who made the creation of this corpus possible by offering their time and support.

7. Bibliographical References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. 2022. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18.

Etienne Barnard, Marelie Davel, and Charl Van Heerden. 2009. Asr corpus design for resource-scarce languages. ISCA.

Etienne Barnard, Marelie H Davel, Charl van Heerden, Febe De Wet, and Jaco Badenhorst. 2014.

[7] The added eight sentences with rare phonemes were put at the end of the list to allow students to gain confidence with the recording process and mitigate the impact of any frustration caused by the rare sounds

[8] https://sadilar.org/en/

The nchlt speech corpus of the south african languages. Workshop Spoken Language Technologies for Under-resourced Languages (SLTU).

Catherine T Best, Gerald W McRoberts, and Elizabeth Goodell. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2):775–794.

Harry Bratt, Leonardo Neumeyer, Elizabeth Shriberg, and Horacio Franco. 1998. Collection and detailed transcription of a speech database for development of language learning technologies. In *ICSLP*.

Noverino N. Canonici. 1989. *Imisindo Yesizulu: A Simple Introduction to Zulu Phonology*. Department of Zulu Language and Literature, University of Natal, Durban.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. `https://github.com/pln-fing-udelar/fast-krippendorff`.

Sylvain Coulange. 2023. Computer-aided pronunciation training in 2022: When pedagogy struggles to catch up. In *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices*, pages 11–22.

Jonathan Dalby and Diane Kewley-Port. 1999. Explicit pronunciation training using automatic speech recognition technology. *CALICO journal*, pages 425–445.

Clement M. Doke. 1961. *Textbook of Zulu Grammar*. Longmans, Cape Town. 6th edition (1st edition 1927).

Olov Engwall. 2012. Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning*, 25(1):37–64.

Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda. 2010. Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401–418.

Sylviane Granger. 2011. How to use foreign and second language learner corpora. *Research methods in second language acquisition: A practical guide*, pages 5–29.

Yakubu A Ibrahim, Juliet C Odiketa, and Tunji S Ibiyemi. 2017. Preprocessing technique in automatic speech recognition for human computer interaction: an overview. *Ann Comput Sci Ser*, 15(1):186–191.

JSM Khumalo. 1984. A preliminary survey of zulu adoptives. *African Studies*, 43(2):205–216.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jing Li, Rui Li, Shen Guo, and Aishan Wumaier. 2023. Enhancing whisper model for pronunciation assessment with multi-adapters. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1955–1959.

Cristóbal Lozano and Amaya Mendikoetxea. 2013. Learner corpora and second language acquisition. *Automatic treatment and analysis of learner corpus data*, 59:65–100.

Brian MacWhinney. 2017. A shared platform for studying second language acquisition. *Language Learning*, 67(S1):254–275.

Roy C Major. 2008. Transfer in second language phonology. *Phonology and second language acquisition*, 36:63–94.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457.

Faria Nazir, Muhammad Nadeem Majeed, Mustansar Ali Ghazanfar, and Muazzam Maqsood. 2023. A computer-aided speech analytics approach for pronunciation feedback using deep feature clustering. *Multimedia Systems*, 29(3):1699–1715.

Hloniphani Ndebele and Nogwaja S Zulu. 2017. The management of isizulu as a language of teaching and learning at the university of kwazulu-natal's college of humanities. *Language and Education*, 31(6):509–525.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. 2021. A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis. In *Interspeech 2021*, page 4448–4452. ISCA.

Nhan Phan, Tamás Grósz, and Mikko Kurimo. 2023. Captaina-a mobile app for practising finnish pronunciation. In *The 24rd Nordic Conference on Computational Linguistics*.

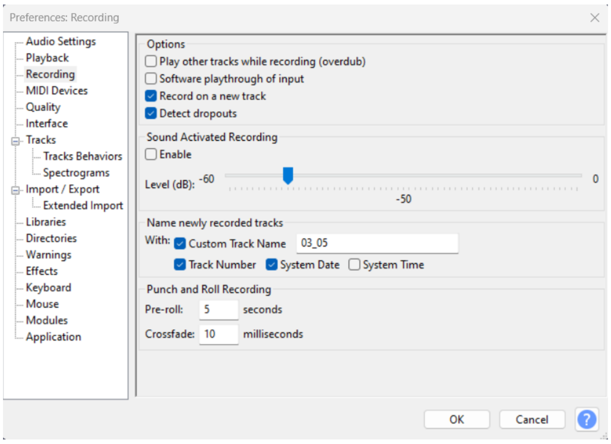
Randi Reppen. 2022. Building a corpus: what are key considerations? In *The Routledge handbook of corpus linguistics*, pages 13–20. Routledge.

Richard W Schmidt. 1990. The role of consciousness in second language learning1. *Applied linguistics*, 11(2):129–158.

M. F. W. 1921. *Elementary Zulu: A Course of Elementary Lessons in the Zulu Language: Intended Chiefly for Beginners and Junior Pupils*. Juta. Google-Books-ID: Wvw0AQAAMAAJ.

Lee Williams. 1979. The modification of speech perception and production in second-language learning. *Perception & Psychophysics*, 26(2):95–104.

A. Audacity Settings



Options

□ Play other tracks while recording (overdub)

□ Software playthrough of input

☑ Record on a new track

☑ Detect dropouts

Sound Activated Recording

□ Enable

Level (dB): -60 0 -50

Name newly recorded tracks

With: ☑ Custom Track Name 03_05

☑ Track Number ☑ System Date □ System Time

Punch and Roll Recording

Pre-roll: 5 seconds

Crossfade: 10 milliseconds

Figure 3: The Audacity settings used for collecting the corpus. The most important setting under the recording options is saving track names automatically, as one can include participant ID and elicitation set information here. In this example, set 3 participant 5. Checking the track number allows data collectors to simply press record and stop repeatedly for each elicited phrase. The exported tracks can then easily and automatically be mapped to the elicitation list, essentially annotating themselves.